

Introduction to Internet of Things
Prof. Sudip Misra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

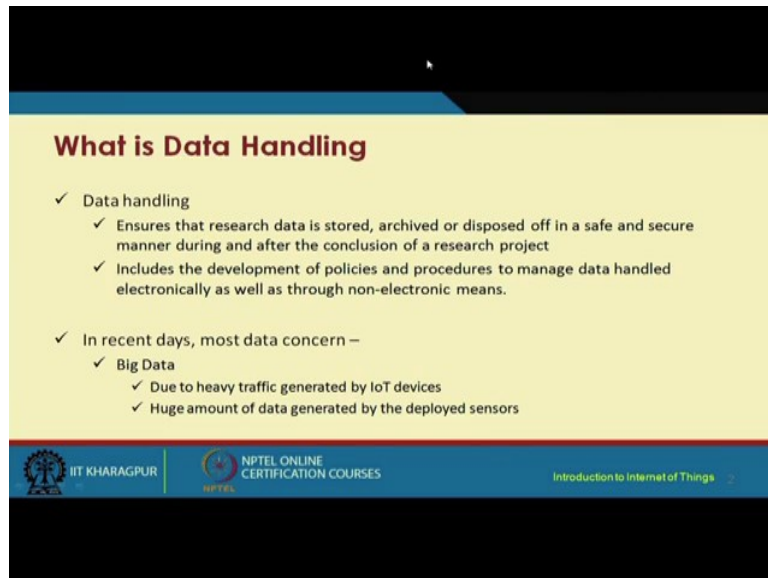
Lecture – 55
Data Handling and Analytics – Part – I

So far, we have understood the different building blocks different technologies about IoT how to build IoT using different technologies we have already gone through in the different lectures. So, we need to now understand that the IoT systems as a whole comprising of devices such as different sensors, actuators and different other communication devices like Wi-Fi 3G, 4G and so on we have mobile devices. So, you know all of these in the IoT world they are huge producers of data. So, IoT is heavily data intensive it is heavily data intensive. So, lot of data gets produced in the IoT deployments in the in the IoT implementation.

So, these data have to be number one properly handled and number 2 is they have to be analyzed to make sense out of the data. So, that things can be made much more efficient where ever the IoT solutions are deployed those problems can be solved can be addressed much more efficiently. So, this is the requirement of the data analytics. So, we have 2 things being discussed in this lecture number one how to handle the data, data handling and number 2 is the data that is generated and is received may be collected at a server either in a centralized way or may it can be distributed the data can be collected in a distributed fashion. So, that data has to be analyzed to make sense out of it to make things better.

So, this is what we are going to look at. So, what are the subtleties what are the important issues concerning this thing. So, this is what we have to understand and this is what we are focusing on in this particular lecture. So, this is divided into data handling and analytics is divided into 2 parts so; however, in both of these lectures what we are going to do is we are simply trying to motivate ourselves and try to understand because this is a this is a introductory course on internet of things here we are not going to understand about the different methodologies for data handling we can or how to how to perform the how to perform the handling of the data or how to analyze the data that we are not going to understand in detail we are simply going to understand that what are the tools the methodology that out there that can be used for handling and analysis of the data.

(Refer Slide Time: 03:27)



What is Data Handling

- ✓ Data handling
 - ✓ Ensures that research data is stored, archived or disposed off in a safe and secure manner during and after the conclusion of a research project
 - ✓ Includes the development of policies and procedures to manage data handled electronically as well as through non-electronic means.
- ✓ In recent days, most data concern –
 - ✓ Big Data
 - ✓ Due to heavy traffic generated by IoT devices
 - ✓ Huge amount of data generated by the deployed sensors

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

So, we start with data handling data handling basically ensures that the data is stored properly, archived properly and disposed off in a safe and secure manner during and after the conclusion of the project. So, I am talking about a project in general in order to understand data handling. So, here we are considered we are considering the development of the policies and the procedures about how to handle the data electronically as well as well as through non electronic means. So, in the IoTs sphere most of the data have certain features which are analogous to the features of big data in other words we are talking about IoT systems producing big data and what is big data that we will understand later, but for now we will just conceive of big data as data that is extraordinarily big in different ways and what are those different ways that we will see later on.

So, due to heavy traffic generated by these IoT devices there is huge amount of data that is created by the different sensors and the different other IoT devices huge amount of data is generated and that data it is big in size continuously big streams of data flow through the network. So, are generated in the network for example, if there is a camera if there is a camera that is fitted. So, that camera basically streams in lot of data continuously. So, in a in a particular hour when the camera stream data is collected you know that will that is a that is huge in size and we are talking about not just one or 2 hours, but we are talking about collecting lot of data over days and months and years and so on.

So, that data has to be at number one stored, number 2 analyzed and then how you are going to handle you know as long as it is not required after that it is no longer required how you are going to dispose of that particular data. So, all these things have to be taken into consideration when we are planning or we are designing an IoT system.

(Refer Slide Time: 06:03)

What is Big Data

- ✓ *"Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis."*
[Report of International Data Corporation (IDC)]
- ✓ *"Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies."*
[National Institute of Standards and Technology (NIST)]

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

So, I was mentioning to you about big data. So, big data there are different there are different definitions of big data. So, one of these definitions talks says that big data technologies describe a new generation of technologies and architectures designed to economically extract value from very large volumes of a very wide variety of data by enabling high velocity capture discovery and or analysis. So, big data shall mean that the data of which the volume acquisition speed or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies.

So, what it means you know all these fancy words have been used in these definitions. So, what it means 2 things data that is huge in size that flows in large velocities that is generated and disseminated in large velocities have to be handled in real time this is one issue, second issue is these data are typically unstructured they are typically unstructured for example, you know text; huge text; Facebook data, Twitter data; you know all these social network data or the data that are generated from the telescopes the sky monitoring telescopes the data that are

generated from you tube and so on. So, these have characteristics which are unstructured they cannot be stored using traditional relational database technologies.

So, how do you handle such data? So, this is a big concern in big data. So, that has to be taken care of. So, you know we cannot we cannot simply be concerned about simply deploying a network without being concerned about how to handle the data that this network is going to produce that is why data handling and big data handling is important.

(Refer Slide Time: 08:17)

Types of Data

- ✓ Structured data
 - ✓ Data that can be easily organized.
 - ✓ Usually stored in relational databases.
 - ✓ Structured Query Language (SQL) manages structured data in databases.
 - ✓ It accounts for only 20% of the total available data today in the world.
- ✓ Unstructured data
 - ✓ Information that do not possess any pre-defined model.
 - ✓ Traditional RDBMSs are unable to process unstructured data.
 - ✓ Enhances the ability to provide better insight to huge datasets.
 - ✓ It accounts for 80% of the total data available today in the world.

The slide is part of an NPTEL online certification course from IIT Kharagpur, titled 'Introduction to Internet of Things'. It features a circular inset image of a man in a white shirt in the bottom right corner.

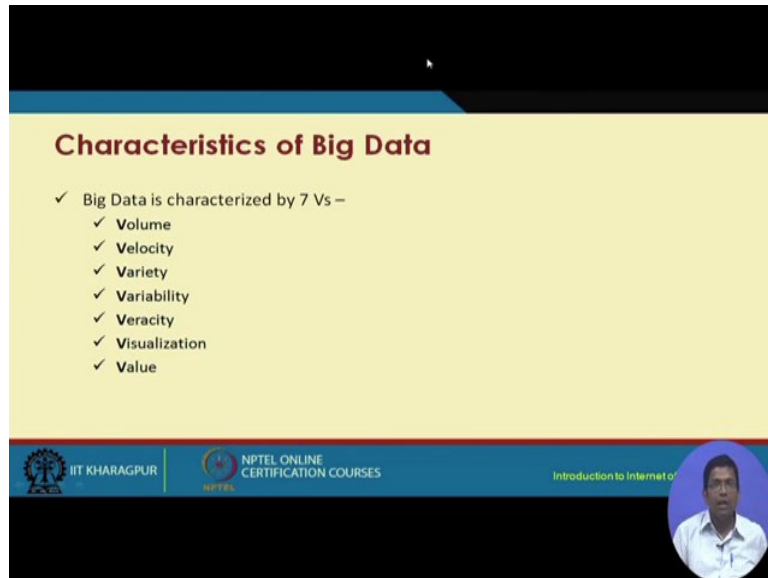
So, we have as I was telling you earlier broadly 2 categories of data structured data like what we have been using library information system student information system accounting information system these are good examples of structured data.

So, these data are easily organized they can be stored in relational databases relational tables you can perform different queries on these data that are stored in the tables and; however, these type of structured data accounts for only twenty percent of total available data in the world today and definitely it is a very small amount of data in the IoT systems. So, IoT systems produce mostly unstructured data which cannot be stored in the form of relational tables.

So, there is not abide by any pre defined relational model for the storing of the data traditional RDBMS techniques are unusable and these data they are very huge in size they are very huge in size and you know most of the data more than 80 percent of the total data that is

available in the world today are in the unstructured form, text fields video, audio speech, you tube data you know telescope data all these are good examples of unstructured data even the data that are produced from most of the IoT devices are unstructured most of the sensor data are unstructured cameras produce unstructured data. So, how do you handle these data?

(Refer Slide Time: 10:09)



Characteristics of Big Data

- ✓ Big Data is characterized by 7 Vs –
 - ✓ Volume
 - ✓ Velocity
 - ✓ Variety
 - ✓ Variability
 - ✓ Veracity
 - ✓ Visualization
 - ✓ Value

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

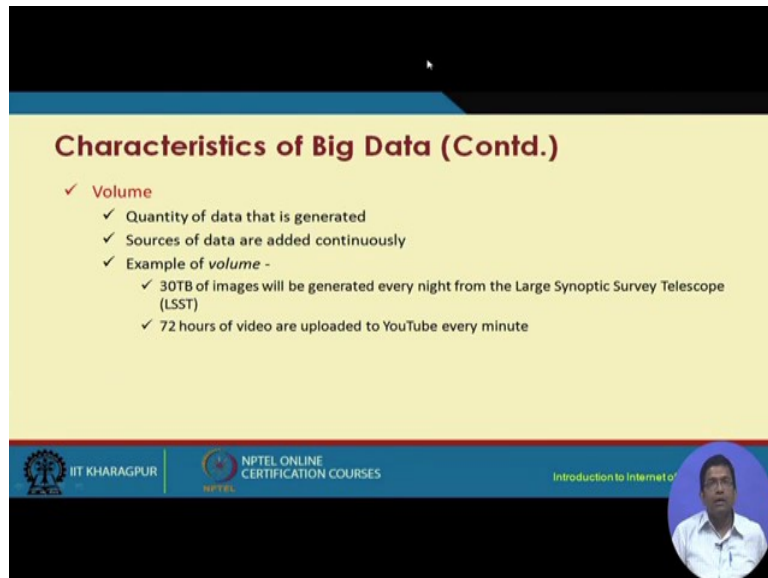
NPTEL

Introduction to Internet of Things

Dr. B. S. Manjunath

So, some of these characteristics of big data earlier it started with a 3 V then came the 5 V definition and now people are talking about 7 V's of big data. So, what are these Vs number 1, volume number 2, V is velocity number 3 is variety, number 4 variability, number 5 veracity, number 6 visualization and number 7 value.


(Refer Slide Time: 10:45)



Characteristics of Big Data (Contd.)

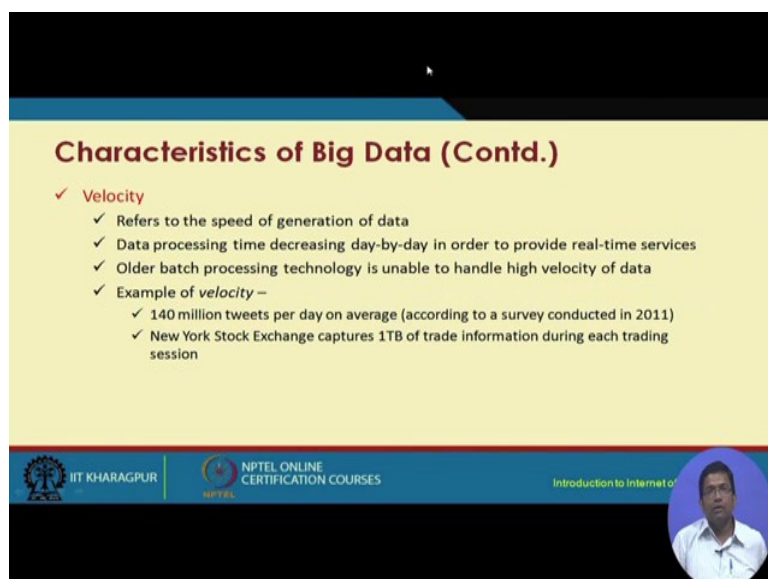
- ✓ **Volume**
 - ✓ Quantity of data that is generated
 - ✓ Sources of data are added continuously
 - ✓ Example of *volume* -
 - ✓ 30TB of images will be generated every night from the Large Synoptic Survey Telescope (LSST)
 - ✓ 72 hours of video are uploaded to YouTube every minute

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



So, these are the 7 different characteristics in to the form of fees of big data. So, let us look at each of them one by one. So, volume; so, big data are characterized with large volumes of data and the quantity of the data that is generated is huge in volume we are talking about more than tera bytes of data several tera bytes of data of images video and so on and so forth you tube you know YouTube basically in every minute you tube in the YouTube 72 hours of video are uploaded. So, it is a huge amount of data huge in terms of volume. So, this has to be in every minute if it is so much. Then just imagine that how much it is going to be every day and in a year how much it is going to be.


(Refer Slide Time: 11:40)



Characteristics of Big Data (Contd.)

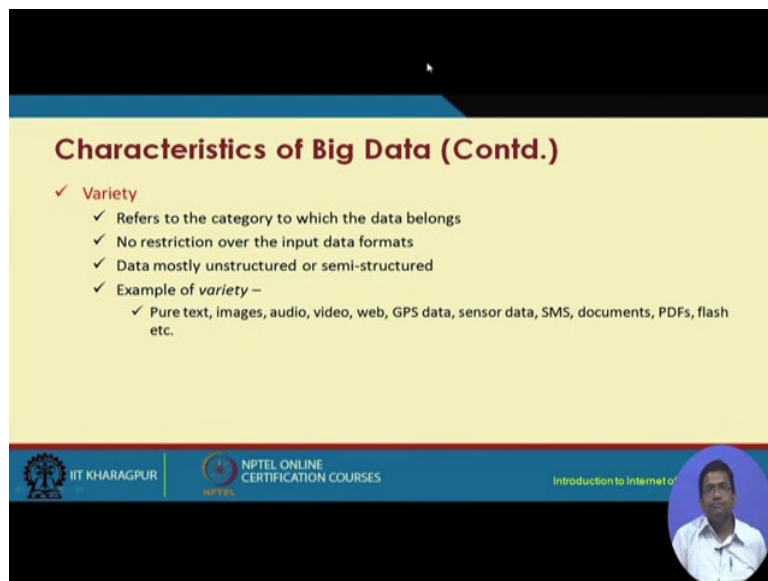
- ✓ **Velocity**
 - ✓ Refers to the speed of generation of data
 - ✓ Data processing time decreasing day-by-day in order to provide real-time services
 - ✓ Older batch processing technology is unable to handle high velocity of data
 - ✓ Example of *velocity* -
 - ✓ 140 million tweets per day on average (according to a survey conducted in 2011)
 - ✓ New York Stock Exchange captures 1TB of trade information during each trading session

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



Velocity as the name suggests; it concerns the speed of generation of the data. So, data processing time is decreasing day by day in order to produce real time services older batch processing technology is unable to handle high velocity of data. So, we need new technologies to handle this high velocity of data. So, these IoT devices mobile phone sensors and so on in huge speeds you know high rate the data is being generated for example, with respect to velocity hundred forty million tweets are generated per day on average the New York stock exchange captures 1 tera byte of trade information during each trading session.

(Refer Slide Time: 12:41)



The slide is titled "Characteristics of Big Data (Contd.)" in a bold, dark red font. Below the title, the word "Variety" is preceded by a red checkmark. Under "Variety", there are four bullet points, each preceded by a black checkmark: "Refers to the category to which the data belongs", "No restriction over the input data formats", "Data mostly unstructured or semi-structured", and "Example of variety –". The last bullet point is followed by a sub-bullet point, also with a black checkmark, listing "Pure text, images, audio, video, web, GPS data, sensor data, SMS, documents, PDFs, flash etc.". The slide has a yellow background for the main content area. At the bottom, there is a blue banner with logos for IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES, and a small circular portrait of a man in a white shirt. The text "Introduction to Internet of Things" is partially visible on the right side of the banner.

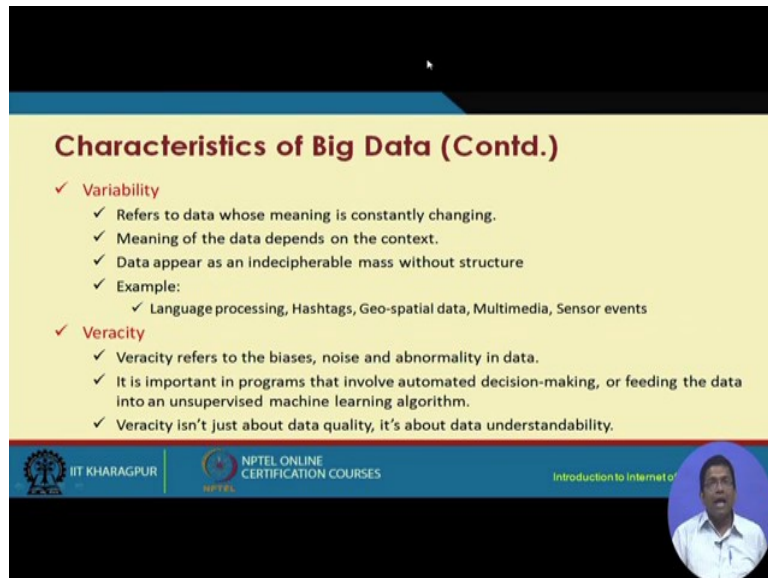
Characteristics of Big Data (Contd.)

- ✓ **Variety**
 - ✓ Refers to the category to which the data belongs
 - ✓ No restriction over the input data formats
 - ✓ Data mostly unstructured or semi-structured
 - ✓ Example of *variety* –
 - ✓ Pure text, images, audio, video, web, GPS data, sensor data, SMS, documents, PDFs, flash etc.

So, just imagine that how much data at what speed is generated in the IoT world.



Variety refers to the category to which the data belongs and most of the data are either unstructured or they are semi structured and examples could be variety could be that pure text data, images, audio, video, web, GPS, then sensor data, SMS, documents, PDFs, flash, etcetera, etcetera. So, all these different varieties of data flowing through a single pipe in the IoT world single pipe huge amounts of data huge volumes of data at high velocities data which is highly varied not only consisting of text, but text audio video images web sensor and so on. So, everything flowing together.


(Refer Slide Time: 13:43)



Characteristics of Big Data (Contd.)

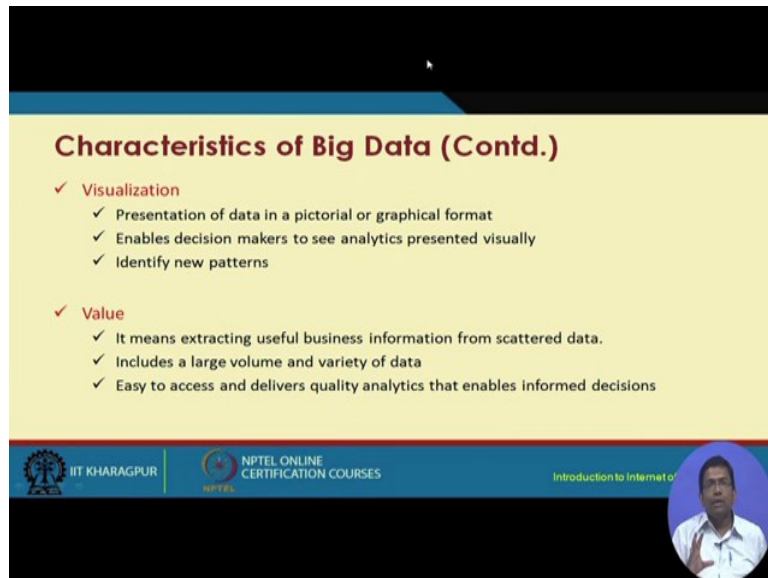
- ✓ **Variability**
 - ✓ Refers to data whose meaning is constantly changing.
 - ✓ Meaning of the data depends on the context.
 - ✓ Data appear as an indecipherable mass without structure
 - ✓ Example:
 - ✓ Language processing, Hashtags, Geo-spatial data, Multimedia, Sensor events
- ✓ **Veracity**
 - ✓ Veracity refers to the biases, noise and abnormality in data.
 - ✓ It is important in programs that involve automated decision-making, or feeding the data into an unsupervised machine learning algorithm.
 - ✓ Veracity isn't just about data quality, it's about data understandability.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



Then we have variability which refers to the data whose meaning is constantly changing the meaning of the data constantly changes depending on the context. So, examples could be language processing. So, language you know it is context dependent. So, you know sometimes language processing it is context driven. So, the meaning basically varies with context hashtags, geo-spatial data, multimedia, sensor events and so on veracity refers to the biases, noise abnormality that exists in the data. So, the IoT data that is typically generated is highly veracious. So, it is important in programs that involve automated decision making or feeding the data into an unsupervised machine learning algorithm. Veracity is not just about data quality it is also about understanding the data.


(Refer Slide Time: 14:44)



Characteristics of Big Data (Contd.)

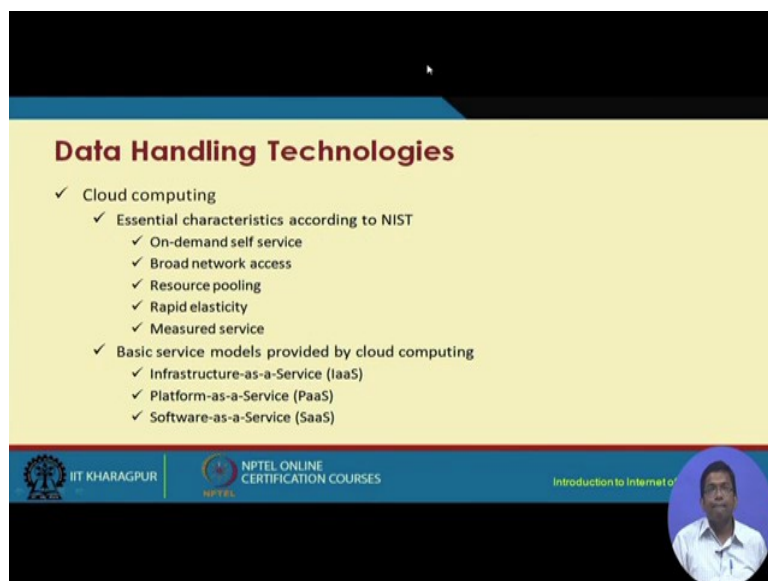
- ✓ **Visualization**
 - ✓ Presentation of data in a pictorial or graphical format
 - ✓ Enables decision makers to see analytics presented visually
 - ✓ Identify new patterns
- ✓ **Value**
 - ✓ It means extracting useful business information from scattered data.
 - ✓ Includes a large volume and variety of data
 - ✓ Easy to access and delivers quality analytics that enables informed decisions

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



Visualization concerns how to present the data pictorially or in a particular easily understandable format it enables the decision makers to see the analytics that are presented visually and identify new patterns value basically means that extracting useful business information from the scattered info. So, how much value the data has from the variety you know. So, from the data; from that how much value it has it includes a large volume and variety of data it is easy to access and deliver quality analytics that enables informed decisions.


(Refer Slide Time: 15:21)



Data Handling Technologies

- ✓ **Cloud computing**
 - ✓ Essential characteristics according to NIST
 - ✓ On-demand self service
 - ✓ Broad network access
 - ✓ Resource pooling
 - ✓ Rapid elasticity
 - ✓ Measured service
 - ✓ Basic service models provided by cloud computing
 - ✓ Infrastructure-as-a-Service (IaaS)
 - ✓ Platform-as-a-Service (PaaS)
 - ✓ Software-as-a-Service (SaaS)

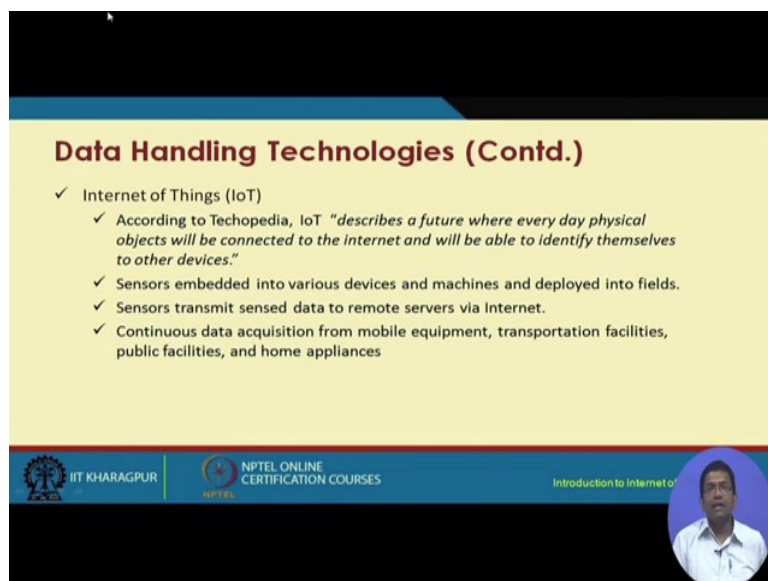
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



There are different data handling technologies that are available for use cloud is one such popular technology. So, here in cloud, cloud basically has some of the essential characteristics as per the definition of NIST on demand self service broad network access resource pooling rapid elasticity measured service means what depending on the amounts of computational resources that are used; it will be built accordingly rapid elasticity means if I need more resources the resources are going to be made available through a pooling based mechanisms. So, resources are going to be pooled from different physical devices and I do not have to own these resources this infrastructure I do not have to own, but I can still get access to these on an on demand manner depending on my requirement and I will build accordingly.

So, some of these basic service models that are there and some of which we have already covered include infrastructure as a service, platform as a service and software as a service.


(Refer Slide Time: 16:51)



Data Handling Technologies (Contd.)

- ✓ Internet of Things (IoT)
 - ✓ According to Techopedia, IoT *"describes a future where every day physical objects will be connected to the Internet and will be able to identify themselves to other devices."*
 - ✓ Sensors embedded into various devices and machines and deployed into fields.
 - ✓ Sensors transmit sensed data to remote servers via Internet.
 - ✓ Continuous data acquisition from mobile equipment, transportation facilities, public facilities, and home appliances

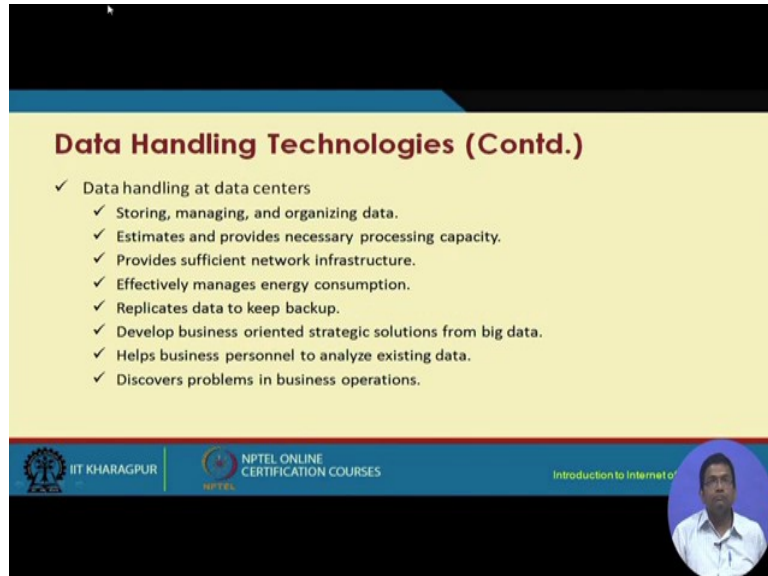
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things | NPTEL



So, cloud is cloud with this 3 basic service models IaaS, PaaS and SaaS is a very important data handling technology that is available to us second is internet of things. So, in the IoT world the sensors that are embedded to the different devices and machines they generate lot of data the sensors transmit this sense data to remote servers via the internet and these data they can be either handled at the back end or these data can also be processed locally at the edge or in the interim somewhere in the intermediate layer. So, continuous data acquisition from mobile equipment transportation facilities public facilities and home appliances are an

important characteristic of IoT and the data that is handled data that is generated in the IoT have to be handled accordingly.


(Refer Slide Time: 17:43)



Data Handling Technologies (Contd.)

- ✓ Data handling at data centers
 - ✓ Storing, managing, and organizing data.
 - ✓ Estimates and provides necessary processing capacity.
 - ✓ Provides sufficient network infrastructure.
 - ✓ Effectively manages energy consumption.
 - ✓ Replicates data to keep backup.
 - ✓ Develop business oriented strategic solutions from big data.
 - ✓ Helps business personnel to analyze existing data.
 - ✓ Discovers problems in business operations.

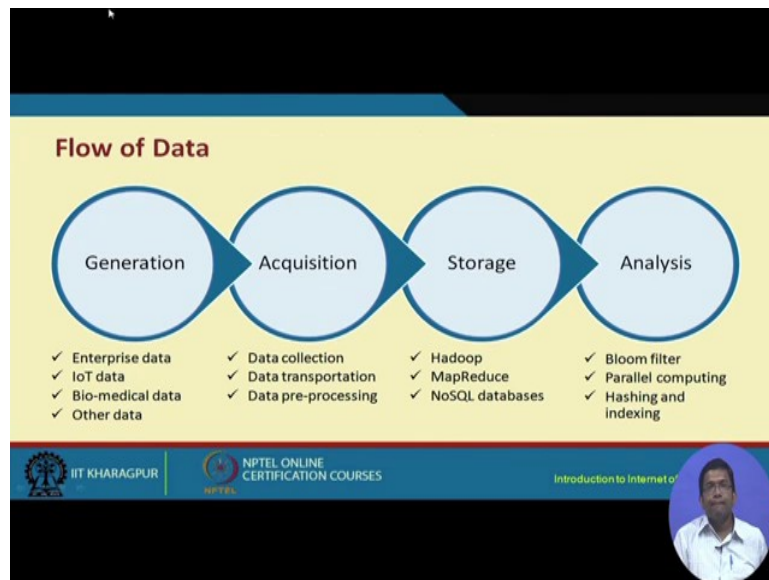
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



Datacenters basically concern you know storing lot of data managing the data organizing the data these data that are generated in the data centers that are that are that exist in the data centers.

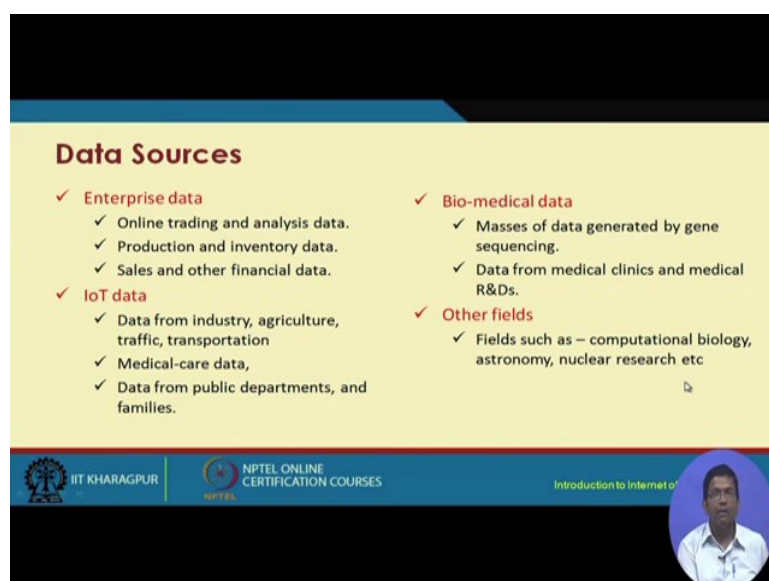
They are they have to be replicated, they have to be backed up they provide sufficient network you know sufficient network infrastructure has to be provided in order to handle this data and this data they can be analyzed in order to discover problems in the business operations.

(Refer Slide Time: 18:25)



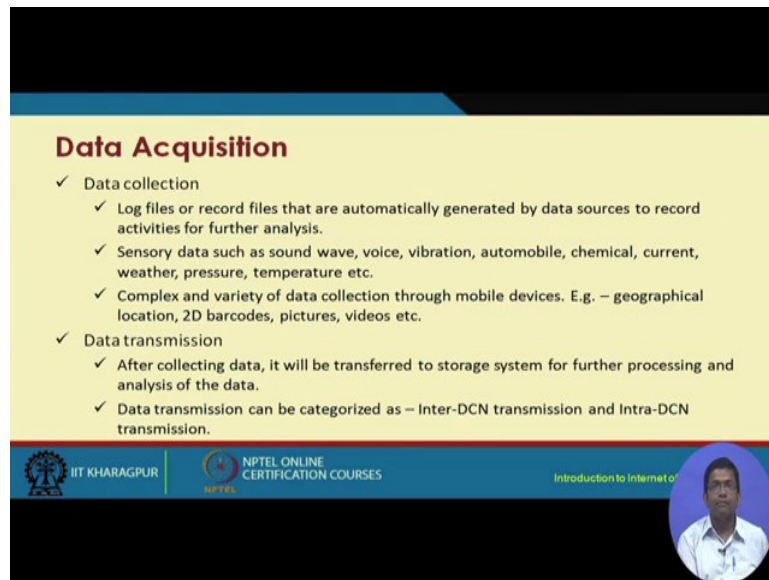
So, this is how the data flows from generation to analysis. So, first the data is generated then comes acquisition of the data storage of the data and finally, analysis of the data. So, in terms of generation of the data from enterprise systems data can be generated from IoT systems from biomedical devices and different other devices all of which are good sources or generators of data in terms of acquisition after the data generation acquisition of the data; data are collected data can be transported data are preprocessed and then data have to be stored we have different technologies for doing it we have Hadoop technology for storage of data MapReduce, NoSQL databases and finally, they have to be analyzed.

(Refer Slide Time: 19:35)



So, for this we have the bloom filter, parallel computing technologies, hashing mechanisms, indexing mechanisms and so on. The different sources of data include enterprise data, IoT data, biomedical data, and other field data from computational biology from nuclear research from astronomy and so on.

(Refer Slide Time: 19:52)



Data Acquisition

- ✓ Data collection
 - ✓ Log files or record files that are automatically generated by data sources to record activities for further analysis.
 - ✓ Sensory data such as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature etc.
 - ✓ Complex and variety of data collection through mobile devices. E.g. – geographical location, 2D barcodes, pictures, videos etc.
- ✓ Data transmission
 - ✓ After collecting data, it will be transferred to storage system for further processing and analysis of the data.
 - ✓ Data transmission can be categorized as – Inter-DCN transmission and Intra-DCN transmission.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things | NPTEL


Now, comes the data acquisition which concerns data collection from log files from different other records activities interviewing if required collecting data from the sensors from the sound sensors voice vibration automobile chemical current weather pressure temperature etcetera and so on. So, after the data are collected they have to be transmitted. So, after the after collecting the data the data have to be transferred to a storage system for further processing and analysis.

(Refer Slide Time: 20:39)

Data Acquisition (Contd.)

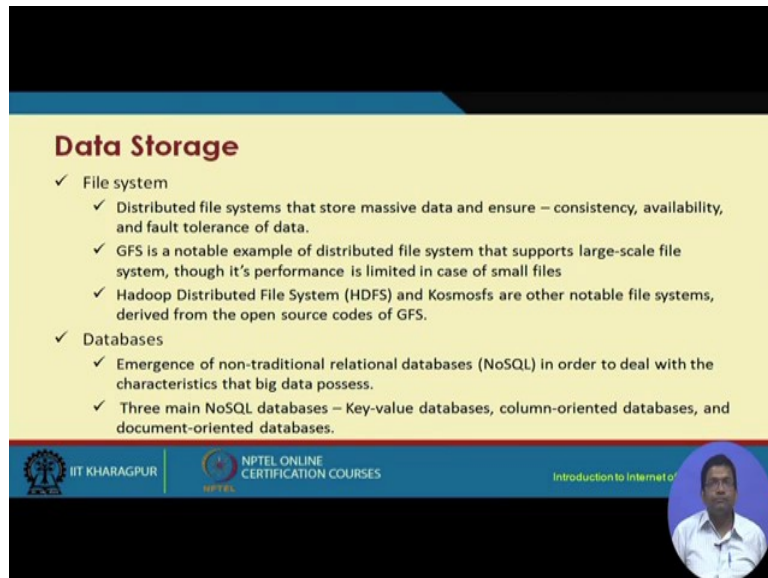
- ✓ Data pre-processing
 - ✓ Collected datasets suffer from noise, redundancy, inconsistency etc., thus, pre-processing of data is necessary.
 - ✓ Pre-processing of relational data mainly follows – integration, cleaning, and redundancy mitigation
 - ✓ Integration is combining data from various sources and provides users with a uniform view of data.
 - ✓ Cleaning is identifying inaccurate, incomplete, or unreasonable data, and then modifying or deleting such data.
 - ✓ Redundancy mitigation is eliminating data repetition through detection, filtering and compression of data to avoid unnecessary transmission.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



So, data transmission can be categorized as in term data center network transmission and intra data center network transmission then the data have to be preprocessed this collected data suffer from noise redundancy inconsistency etcetera. So, these have to be removed they have to be preprocessed the data have to be preprocessed the preprocessing of the relational data mainly follows integration cleaning and redundancy mitigation integration is combining the data from various sources and providing users with a information with uniform view of the data cleaning of the data is required in order to remove inaccuracies, incompleteness and unreasonable behavior of the data unreasonable characteristics of the data and then either modifying the data or to remove these problems or deleting these data all together.


(Refer Slide Time: 21:31)



Data Storage

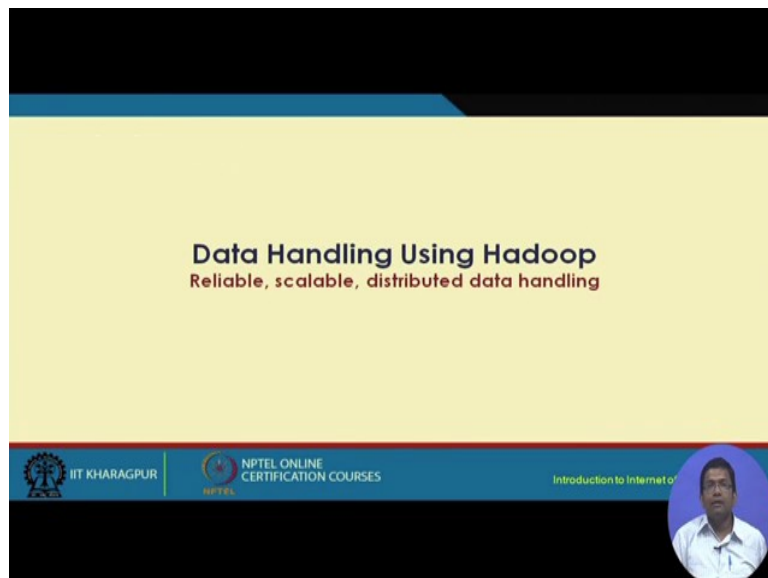
- ✓ File system
 - ✓ Distributed file systems that store massive data and ensure – consistency, availability, and fault tolerance of data.
 - ✓ GFS is a notable example of distributed file system that supports large-scale file system, though its performance is limited in case of small files
 - ✓ Hadoop Distributed File System (HDFS) and Cosmosfs are other notable file systems, derived from the open source codes of GFS.
- ✓ Databases
 - ✓ Emergence of non-traditional relational databases (NoSQL) in order to deal with the characteristics that big data possess.
 - ✓ Three main NoSQL databases – Key-value databases, column-oriented databases, and document-oriented databases.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things




Now, the data after acquisition the data has to be stored the data can be stored in the file systems or in data bases. So, if we are talking about relational databases SQL is good enough; however, with the kind of data that is exhibited then unstructured data that is exhibited these NoSQL is very useful. NoSQL basically uses 3 different types of databases one is the key value database the second is the column oriented database and the third is the document oriented data base.

(Refer Slide Time: 22:18)



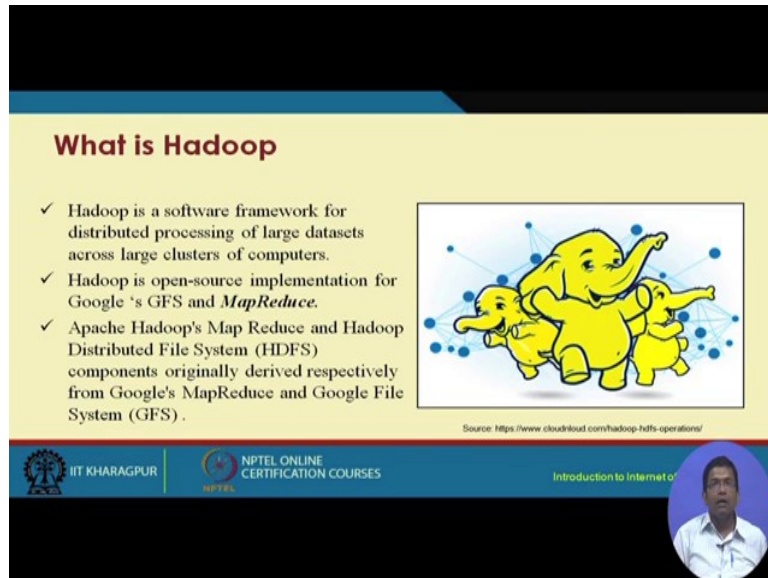
Data Handling Using Hadoop
Reliable, scalable, distributed data handling

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



So, this is about the data handling the different aspects of data handling for this what technology do we have for handling this kind of data what technology do we have we have the Hadoop technology.

(Refer Slide Time: 22:36)




What is Hadoop

- ✓ Hadoop is a software framework for distributed processing of large datasets across large clusters of computers.
- ✓ Hadoop is open-source implementation for Google's GFS and *MapReduce*.
- ✓ Apache Hadoop's Map Reduce and Hadoop Distributed File System (HDFS) components originally derived respectively from Google's MapReduce and Google File System (GFS).

Source: <https://www.cloudnexus.com/hadoop-hdfs-operations/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things



So, what is Hadoop this is basically a software framework for distributed processing of large datasets across large clusters of computers it is a open source implementation for Google's GFS and MapReduce. GFS is basically the Google file system and MapReduce apaches apache Hadoops MapReduce and Hadoop distributed file system which in short is called is well known as HDFs has different components which are originally derived respectively from GFS, Google's, MapReduce and GFS full file system.

(Refer Slide Time: 23:20)

Building Blocks of Hadoop

- ✓ Hadoop Common
 - ✓ A module containing the utilities that support the other Hadoop components
- ✓ Hadoop Distributed File System (HDFS)
 - ✓ Provides reliable data storage and access across the nodes
- ✓ MapReduce
 - ✓ Framework for applications that process large amount of datasets in parallel.
- ✓ Yet Another Resource Negotiator (YARN)
 - ✓ Next-generation MapReduce, which assigns CPU, memory and storage to applications running on a Hadoop cluster.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

So, these are building blocks of Hadoop; Hadoop common HDFS is Hadoop distributed file system MapReduce and YARN which stands for yet another resource negotiator so, without going through each of these in further detail.

(Refer Slide Time: 23:41)

Hadoop Distributed File System (HDFS)

- ✓ Centralized node
 - ✓ Namenode
 - ✓ Maintains metadata info about files
- ✓ Distributed node
 - ✓ Datanode
 - ✓ Store the actual data
 - ✓ Files are divided into blocks
 - ✓ Each block is replicated

Block Replication

Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, (1,3), ...
/users/sameerp/data/part-1, r:3, (2,4,5), ...

Datanodes

Diagram showing 8 Datanodes arranged in a 2x4 grid. Each Datanode contains a set of blocks (represented by numbers in boxes). The blocks are replicated across the Datanodes. For example, block 1 is replicated on Datanodes 1, 2, and 3. Block 2 is replicated on Datanodes 2, 4, and 5. Block 3 is replicated on Datanodes 3, 6, and 7. Block 4 is replicated on Datanodes 4, 7, and 8. Block 5 is replicated on Datanodes 5, 8, and 6.

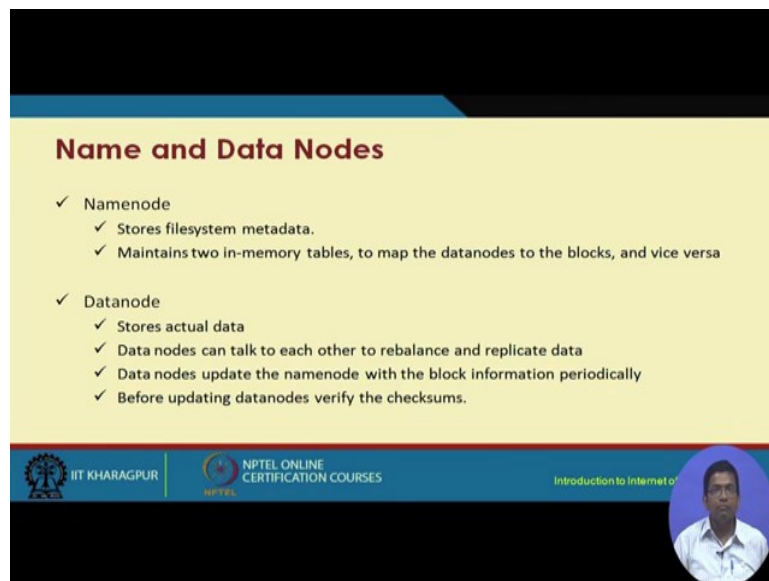
Source: http://hadoop.apache.org/docs/1.2.1/hdfs_design.html

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

We will just look at only HDFS the Hadoop distributed file system. So, in HDFS which is which is basically the important thing in Hadoop. So, in HDFS there are 2 different types of nodes we have the name node and we have the data node. So, the name node is a centralized node which maintains the metadata information about the different files storing the data and

data node is a distributed node that stores the actual data in the form of files which are again divided into blocks and each of these blocks is replicated and this is what is shown over here in this particular figure. So, what you see over here are these data nodes with replication of the blocks and this is the name node. So, name node has the metadata and the data nodes have the actual data and these data nodes are fragmented into blocks and these blocks are replicated and consequently we have more reliability in the storage of the data in Hadoop HDFS.


(Refer Slide Time: 24:57)



Name and Data Nodes

- ✓ Namenode
 - ✓ Stores filesystem metadata.
 - ✓ Maintains two in-memory tables, to map the datanodes to the blocks, and vice versa
- ✓ Datanode
 - ✓ Stores actual data
 - ✓ Data nodes can talk to each other to rebalance and replicate data
 - ✓ Data nodes update the namenode with the block information periodically
 - ✓ Before updating datanodes verify the checksums.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

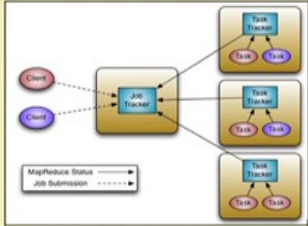


So, we have the name node which stores the file system meta data and the data node which stores the actual data.

(Refer Slide Time: 25:06)

Job and Task Trackers

- ✓ Job Tracker –
 - ✓ Runs with the Namenode
 - ✓ Receives the user's job
 - ✓ Decides on how many tasks will run (number of mappers)
 - ✓ Decides on where to run each mapper (concept of locality)
- ✓ Task Tracker –
 - ✓ Runs on each datanode
 - ✓ Receives the task from Job Tracker
 - ✓ Always in communication with the Job Tracker reporting progress



The diagram illustrates the interaction between a Job Tracker and multiple Task Trackers. On the left, a 'Client' (represented by a red circle) sends a 'MapReduce Status' (solid arrow) and a 'Job Submission' (dashed arrow) to the 'Job Tracker' (a central box). The 'Job Tracker' then distributes tasks to three 'Task Tracker' boxes on the right. Each 'Task Tracker' contains two 'Task' boxes. A legend at the bottom left of the diagram shows a solid arrow for 'MapReduce Status' and a dashed arrow for 'Job Submission'. Source: <http://developing.in/articles/2015/aug/11/an-introduction-to-apache-hadoop-for-big-data/>

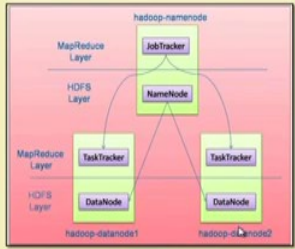
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

Now, we also have these job and job trackers. The job tracker basically runs with the name node it receives the users job and decides on how how many tasks will run and the job tracker basically runs on each data node receives the task from the job tracker and it is always in communication with the job tracker reporting it the progress that is made. So, as we can see over here we have the job tracker and these different tasks which are monitored through the task tracker. So, we have the job tracker running on the name node and the task tracker which is running on the data nodes.

(Refer Slide Time: 25:57)

Hadoop Master/Slave Architecture

- ✓ Master-slave *shared-nothing* architecture
- ✓ Master
 - ✓ Executes operations like opening, closing, and renaming files and directories.
 - ✓ Determines the mapping of blocks to Datanodes.
- ✓ Slave
 - ✓ Serves read and write requests from the file system's clients.
 - ✓ Performs block creation, deletion, and replication as instructed by the Namenode.



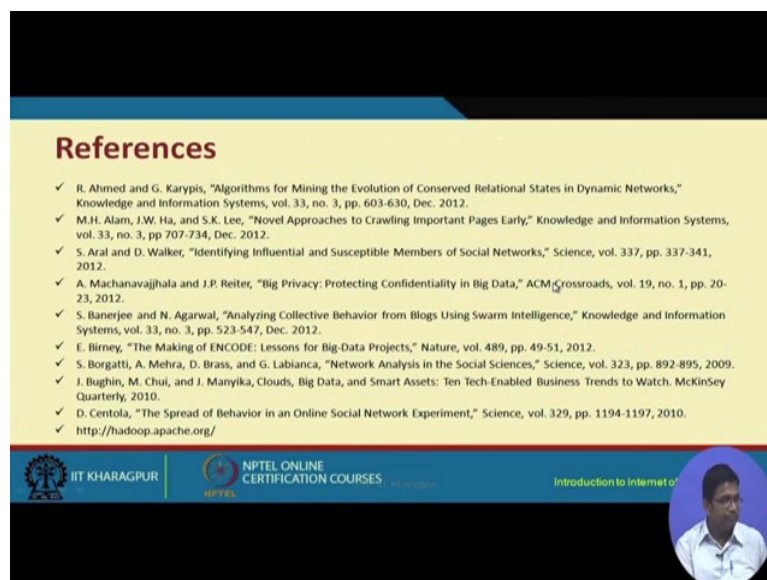
The diagram shows the Hadoop Master/Slave Architecture. It is divided into three layers: MapReduce Layer, HDFS Layer, and another MapReduce Layer. In the top MapReduce Layer, there is a 'JobTracker' box. In the HDFS Layer, there is a 'NameNode' box. In the bottom MapReduce Layer, there are two 'TaskTracker' boxes. In the bottom HDFS Layer, there are two 'DataNode' boxes. The 'JobTracker' is connected to the 'NameNode' and the two 'TaskTracker' boxes. The 'NameNode' is connected to the two 'DataNode' boxes. The 'TaskTracker' boxes are connected to the 'DataNode' boxes. Source: <http://ankitashblogger.blogspot.in/2011/01/hadoop-cluster-setup.html>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Introduction to Internet of Things

So, here what we see is the master slave architecture in Hadoop. So, we have this kind of thing we have the name node like this which contacts the metadata, the name node and then we have these different data nodes. This name node has the job tracker and the name node information and this name node basically points to the data node of the data node in the in the data node that is installed right and then we have the MapReduce.

So, the map; so, what we have the task the task tracker is basically linked with the job tracker job tracker and the task tracker. So, job tracker in the name node is related to the; is linked with the task tracker in the data node. So, the job tracker decides in the MapReduce layer the name node in the HDFS layer task tracker in the MapReduce layer data node in the HDFS layer and so on. So, this is the architecture the master slave architecture in Hadoop.

(Refer Slide Time: 27:17)



So, with this we come to an end of the discussions on the file handling and particularly focusing on file handling of this lecture on file handling and data analysis and here some of these references are there and with this we come to an end. And so, what we have discussed is how data handling is important what are the different sources of data how data have to be handled and how Hadoop and its different components can come as an aid for handling data which is which has the features of big data data that is generated from the IoT systems how it can be how Hadoop can be used in order to handle this kind of data.

Thank you.